

ПОПОЛНЕНИЕ ЯЗЫКОВОЙ БАЗЫ ЗНАНИЙ В ЗАДАЧЕ АНАЛИЗА ЭКВИВАЛЕНТНОСТИ СМЫСЛОВЫХ ОБРАЗОВ ВЫСКАЗЫВАНИЙ.

Г.М.Емельянов, Д.В.Михайлов, Н.А.Степанова

Новгородский государственный университет имени Ярослава Мудрого

Ряд задач семантического анализа высказываний на Естественном Языке (ЕЯ) заключается во взаимном сопоставлении смыслов на предмет тождественности (эквивалентности). В частном случае некоторый смысл принимается за эталон и ставится задача оценить степень близости смысла высказывания заданному “правильному” смыслу. Наиболее актуальной на сегодняшний день практической задачей, требующей сравнение смысла высказывания с эталоном, является интерпретация тестовых заданий открытой формы в системах автоматизированного тестирования и контроля знаний. Тестовые задания открытой формы требуют от обучаемого формулирования развернутого ответа на поставленный системой вопрос (*плакат 2*). Как показывает опыт разработки различных тестовых систем, применение открытых тестов затруднено в силу ряда причин. Одна из них заключается в необходимости оперирования большим количеством сущностей при интерпретации теста и, как следствие, отсутствие универсальных механизмов оценки правильности ответа. Наиболее разумным путем решения указанной проблемы является именно введение “эталонного” смысла, относительно которого ведется сравнение (*плакаты 3 и 4*).

Использование языка глубинного синтаксиса в качестве языка смыслов в рамках теоретического подхода к языку как преобразователю “Смысл \Leftrightarrow Текст” дает возможность решить задачу оценки близости смысла анализируемого высказывания (ответа на вопрос теста) эталону посредством конечного числа корректно формализуемых правил синонимических преобразований помеченных деревьев (*плакат 6*). Указанные правила ЛФ-синонимических преобразований описывают ситуации лексико-синтаксических замен на уровне варьирования универсальной (абстрактной) лексики в рамках аппарата стандартных Лексических Функций (ЛФ), что особенно актуально для реальных тестов : в большинстве случаев обучаемый употребляет синонимы именно на уровне абстрактных слов и их сочетаний, оставляя предметную лексику без изменений (*плакаты 5 и 7*).

В представленной работе исследуется ряд практических аспектов реализации механизма установления смысловой эквивалентности (*плакат 1*). Следует отметить, что построение дерева глубинного синтаксиса как подлежащего анализу формального образа смысла высказывания требует последовательного выполнения морфологического и синтаксического анализа высказывания (*плакат 8*). При этом для каждого предложения в тексте строится дерево зависимостей (поверхностная синтаксическая структура), которая затем преобразуется в глубинную синтаксическую структуру с привлечением информации :

- Моделей управления (*плакат 9*) - в соответствии с их описанием по Толково-комбинаторному словарю – для идентификации глубинных синтаксических актантов слов;
- Базы данных лексических функций – с целью выявления лексических коррелятов самостоятельных лексем в дереве глубинного синтаксиса.

Технологии морфологического и синтаксического анализа текстов на сегодняшний день проработаны в достаточной степени и имеют практическое применение в текстовых процессорах, в том числе MS Word, в то время как задачу полностью автоматизированного глубинного синтаксического анализа нельзя считать решенной даже в первом приближении. Для корректной работы алгоритма считывания глубинной синтаксической структуры с дерева синтаксического подчинения необходимо иметь формализованное описание моделей управления всех лексем используемого подмножества ЕЯ, с которыми могут находиться в отношении подчинения другие лексемы. Указанное требование относится как к абстрактной универсальной, так и предметной лексике. Описания моделей управления абстрактных лексем могут быть заложены в базу изначально, при построении системы. Предметная лексика описывается в процессе настройки на конкретную область знаний, а также в процессе эксплуатации системы. Построение моделей управления для новых слов предполагает наличие у сопровождающих систему специалистов лингвистических навыков. Поэтому актуальной здесь является автоматизация построения моделей управления.

Решение указанной проблемы предполагает решение следующих основных *задач* :

- Разработка и исследование методов автоматического пополнения словарей.
- Разработка структуры языковой базы знаний для всех указанных на *плакате 2* этапов обработки входного текста с учетом возможности ее автоматического пополнения;

Для решения задачи пополнения Базы Знаний Моделей Управления авторами настоящей публикации первоначально было решено использовать компьютерный тезаурус RussNet – Wordnet-тезаурус русского языка, разрабатываемый исследовательским коллективом кафедры математической лингвистики Санкт-Петербургского государственного университета.

В настоящее время Wordnet-тезаурусы является одним из самых распространенных типов лингвистических ресурсов в сфере информационных технологий. Данный тип компьютерного тезауруса являются удобным формализмом для представления структуры словарного состава именно специальных предметно-ориентированных подмножеств естественного языка.

Валентностные фреймы представленных в RussNet глаголов могут быть напрямую отображены в описания способов поверхностной реализации и семантических интерпретаций глубинных синтаксических актантов (мест Модели Управления).

В частности (*плакаты 11, 12, 13*) :

- нумерация валентностей глагола соответствует обозначениям типов отношения подчинения между лексемой и ее глубинным синтаксическим актантом;
- тематические роли глагольных аргументов соответствуют ролям обозначаемых глагольными актантами сущностей;
- базовый концепт глагольного аргумента соответствует семантическому классу рассматриваемой аргументом сущности (по Модели Управления);
- информация о наличии/отсутствии предлога, обязательной в поверхностной реализации глагольного аргумента, символическое обозначение синтаксического класса глагольного аргумента и числовой код грамматической информации являются частью описания способа поверхностной реализации глубинного синтаксического актанта лексемы.

Тем не менее, на сегодняшний день объем содержащейся в RussNet лингвистической информации недостаточен для описания всех мыслимых в тестировании подмножеств русского языка, относящихся к разным областям знаний. Система тестирования знаний, в принципе, строится намного быстрее, чем словарное описание ЕЯ, необходимое для ее успешного функционирования. В этой ситуации время построения тезауруса оказывается критичным фактором. Наиболее разумным путем решения данной проблемы является разработка тезаурусов “типа RussNet” рядом исследовательских коллективов параллельно, но при этом основная идея RussNet должна строго выдерживаться. Данный путь решения немыслим без автоматизации труда лексикографа.

Отличительной особенностью RussNet является описание характерных для русского языка семантических деривационных связей между глаголами и их производными. Глагол-гипоним полностью наследует валентностную структуру своего гиперонима. Представленные в RussNet отглагольные существительные, выражающие действия, наследуют аргументную структуру глагола, включая сочетаемостные ограничения. Пример : *бороться за правое дело – борьба за правое дело*. Указанное свойство wordnet-описания лексической системы позволяет путем использования механизма наследования определять структуру моделей управления существительных, которые являются формальными или семантическими производными от глаголов.

Для логического моделирования и исследования свойств лексической системы ЕЯ, в частности, при описании ее wordnet-тезаурусами, для решения ряда прикладных задач за рубежом широко используется подход на основе Формального Концептуального Анализа (ФКА).

ФКА – это метод анализа данных, основанный на математической теории решеток. Впервые теория ФКА была представлена в 1980 году исследовательской группой под руководством Рудольфа Вилле (Rudolf Wille) из Darmstadt University, Германия. Начиная с этого времени, в Западной Европе, Австралии, США по этой проблематике было опубликовано несколько сотен статей, фундаментальные математические исследования и курсы лекций. Изначально ФКА развивался как математическая теория, ставившая своей целью реструктуризацию классической теории решеток для ее прагматического

применения. Однако последние годы основное направление исследований перешло из области чистой математики в область прикладной математики и информатики. В настоящее время ФКА используется как математическая основа Концептуальной Обработки Знаний (Conceptual Knowledge Processing), теории имеющей своей целью представить методы и инструменты для человеко-ориентированной концептуальной обработки знаний. Формальный Концептуальный Анализ это (плакат 14) :

- математическое описание философского понятия концепта;
- ориентированный на человека метод структурирования и анализа данных;
- метод для визуализации данных, представления иерархий и зависимостей.

Приведем математическое описание ФКА.

- 1) Концепт и формальный контекст (плакат 15). Концепт: объекты (G), атрибуты (M) и отношения между ними (I). Формальный контекст: это тройка (G, M, I) , может быть задан таблицей.
- 2) Формальный концепт (плакат 16). Деривационный оператор применяется к как множеству объектов, так и к множеству атрибутов :

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$

$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

Формальный концепт, определенный для формального контекста (G, M, I) , это пара (A, B) , где $A \subseteq G$, $B \subseteq M$, $A' = B$ и $B' = A$. Множество A называют степенью (extent) концепта, а B – целью (intent) концепта. Множество всех формальных концептов для контекста (G, M, I) обозначается $\beta(G, M, I)$.

- 3) Концептуальная решетка (плакат 17). Наиболее важная структура, определенная на $\beta(G, M, I)$ это формальное отношение субконцепт-суперконцепт: формальный концепт $c1$ является субконцептом концепта $c2$ $(A1, B1) \leq (A2, B2)$, если $A1 \subseteq A2$ или $B2 \subseteq B1$, и $c2$ является формальным суперконцептом $c1$ $(A2, B2) \leq (A1, B1)$, если $A2 \subseteq A1$ или $B1 \subseteq B2$. Множество $\beta(G, M, I)$ всех концептов формального контекста с определенным отношением порядка (\leq) называется концептуальной решеткой (плакат 18).

Формальный Концептуальный Анализ является естественным способом представлением иерархий и классификации по атрибутивным признакам. Иерархические отношения являются центральной частью любой лексической базы данных. ФКА может быть использован для моделирования следующих отношений в тезаурусе RussNet: гипонимия (родовидовое) и меронимия (часть-целое). В работе предлагается использовать лексемы в качестве объектов, а в качестве атрибутов могут быть использованы элементы толкования слова, синтаксические и семантические признаки глаголов, а также структура и элементы валентностных фреймов (плакат 19). В результате генерируется концептуальная решетка (плакат 20), где формальные концепты соответствуют синсетам тезауруса RussNet, а отношение порядка моделирует иерархические отношения тезауруса. В итоге концептуальная решетка позволяет представить тезаурус в избыточном представлении, облегчить поиск лексем и

использовать валентностные фреймы лексем, со схожими атрибутами, в качестве шаблонов для моделей управления.

При использовании синтаксических и семантических валентностей слова, соответствующие глубинным синтаксическим актантам, в качестве определительных атрибутов лексического значения слова встает вопрос об однозначной идентификации самих этих атрибутов.

Для решения указанной проблемы, а также возможности автоматического построения описания каждого актанта слова авторами настоящей публикации предлагается использовать сформулированные академиком Ю.Д. Апресяном выводы о :

- зависимости ролевого состава семантических валентностей слова от того семантического класса, к которому данное слово принадлежит;
- наличию типовых форм выражения актантных значений;
- зависимости типичных синтаксических функций словоформ, реализующих валентности, (номеров глубинных синтаксических актантов, типов отношения глубинного синтаксиса между словом и актантом) от естественного порядка появления в тексте этих словоформ.

Имея (*плакат 21*) реализованные в виде утверждений динамической БД справочники :

- семантических ориентаций (каждому слову ставится в соответствие его семантический класс);
- ролевого состава лексических значений (лексике заданного семантического класса ставится в соответствие список ролей потенциальных актантов слова);
- типовых форм выражения ролей (заданной роли ставится в соответствие семантическая ориентация актанта, синтаксический класс, код грамматической информации, обязательный в поверхностной реализации актанта предлог и типичная лексика для обозначения роли),

можно строить описания актантов нового слова в соответствии с представленным на *плакате 21* ориентировочным Пролог-правилом. При этом пользователю достаточно иметь минимальные лексикографические навыки.